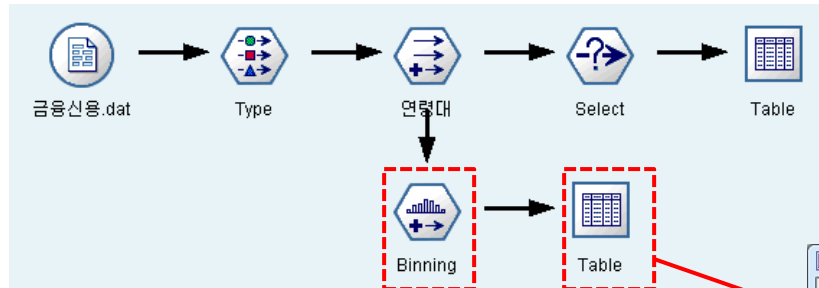




Data Munging with SPSS Modeler

데이터 구간화



❖ 구간화의 필요성

- 연속형 데이터의 범주화
- 범주화를 통해서 정확한 의미 부여
- 각종 분위수의 산정을 통한 정보의 이해 증진

❖ 구간화(Binning) 노드

연속형 변수를 등간격, 등비율, 평균/표준편차 등을 고려한 형태의 구간화 규칙을 적용하여 범주형 변수로 변환

Binning Dialog Box Settings:

- Bin fields: 월평균수입(1000원)
- Binning method: Tiles (equal count) ✓
- Equal Count Binning:
 - Tile name extension: _TILE
 - Custom tile extension: _TILEN
 - ☒ Quartile (4)
 - ☒ Decile (10)

Callout Bubble: 월평균수입 변수를 4 분위수와 10분위수 두가지로 구간화함

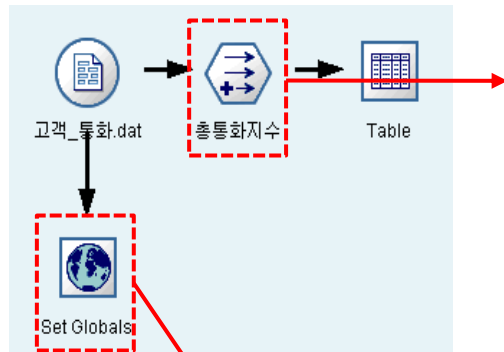
Data Table (Table 15 fields, 357 records):

미 된 년수	현 직장에서 근무한 년수	은행에서 판별한 신용도 정도	연령대	월평균수입(1000원)_TILE4	월평균수입(1000원)_TILE10
1	5	5 양호	20대이하	2	3
2	3	6 모호	30대	4	9
3	7	2 모호	20대이하	2	3
4	8	9 양호	40대이상	4	9
5	10	10 양호	40대이상	4	10
6	3	2 양호	20대이하	1	3
7	4	4 양호	40대이상	3	6
8	3	4 양호	30대	3	6
9	4	5 양호	40대이상	3	7
10	4	8 양호	30대	3	6
11	6	8 모호	20대이하	4	9

전역 값 설정

실습 DATA

- 고객_통화.dat



Derive

Derive as: Formula

Mode: ☒ Single ☐ Multiple

Derive field:

총통화지수

Derive as: Formula

Field type: <Default>

Formula:

총통화시간_분 / @GLOBAL_MEAN(총통화시간_분)

전역값(Filler) 노드

: 필드의 통합 계산된 값을 memory에 저장하여 스트림내에서 자유롭게 사용할 수 있게 한다.

설정한 전역값을 파생 노드를 통해 새로운 변수 생성에 사용함

Set Globals

Globals to be created:

Field	MEAN	SUM	MIN	MAX	SDEV
총통화시간_분	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Default operation(s): ☒ MEAN ☒ SUM ☒ MIN ☒ MAX ☒ SDEV

☒ Clear all globals before executing

☐ Display preview of globals created after execution

Settings Annotations

OK Execute Cancel Apply Reset

총통화시간_분의 전역값을 설정함

Table (26 fields, 3,155 records) #1

	회선	국내통화시간	평균국내통화시간	총통화시간	총통화지수
1	88	152.100	1.728	176.640	0.148
2	56	119.758	2.139	145.538	0.122
3	50	147.600	2.952	149.568	0.125
4	38	126.600	3.332	140.820	0.118
5	35	79.029	2.258	84.090	0.070
6	24	74.222	3.093	74.342	0.062
7	38	126.600	3.332	140.820	0.118
8	18	175.500	9.750	185.350	0.155
9	15	128.100	8.540	168.100	0.140
10	45	64.400	1.431	72.431	0.061
11	0	0.000	0.000	0.000	0.000
12	66	87.450	1.325	106.480	0.089
13	0	0.000	0.000	0.000	0.000
14	85	88.092	1.036	103.044	0.086
15	78	169.200	2.169	221.342	0.185
16	25	72.311	2.892	80.612	0.067
17	1	1.585	1.585	11.511	0.010
18	54	124.900	2.313	150.038	0.125
19	43	157.500	3.663	208.558	0.174
20	65	113.488	1.746	114.320	0.096

Table Annotations

OK

결측 값 처리 (1/2)

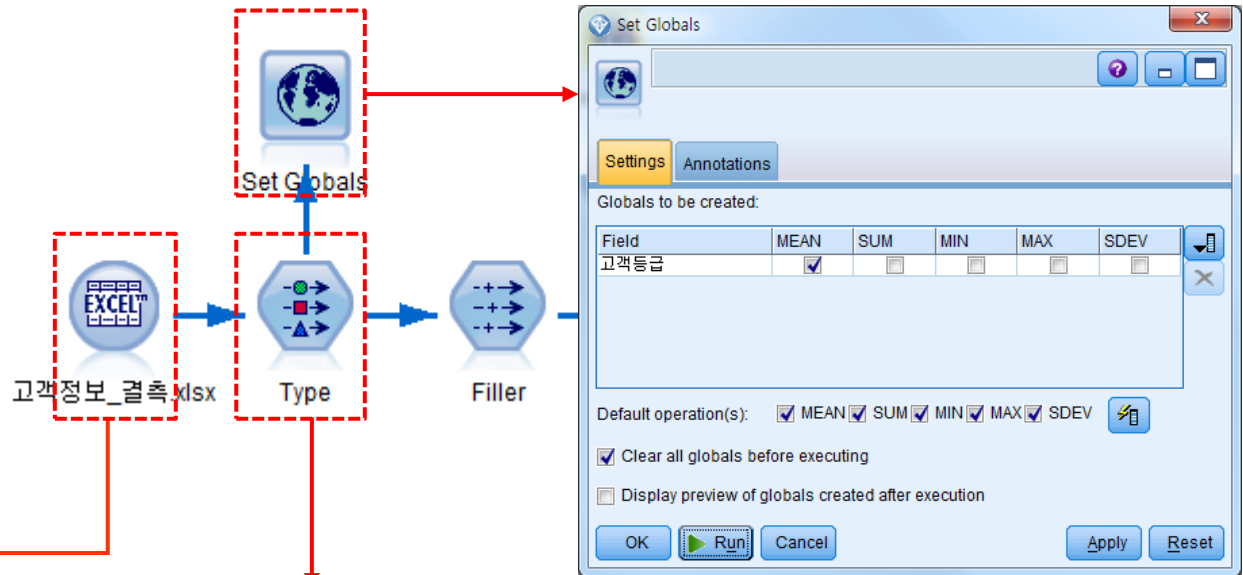
실습 DATA

- 고객정보_결측.xlsx

결측값이 있는 data

Table (11 fields, 31,7...)

	이탈여부	핸드셋	고객등급
41	이탈	SOP20	1.000
42	이탈	SOP10	4.000
43	이탈	SOP10	4.000
44	이탈	SOP20	1.000
45	이탈	SOP20	5.000
46	이탈	SOP10	3.000
47	이탈	SOP10	\$null\$
48	이탈	SOP10	5.000
49	이탈	SOP20	2.000
50	유지	SOP20	4.000
51	이탈	SOP20	5.000
52	이탈	SOP10	4.000
53	이탈	SOP20	3.000
54	이탈	SOP10	1.000
55	이탈	SOP10	2.000
56	이탈	SOP20	5.000
57	이탈	SOP10	\$null\$
58	이탈	SOP20	3.000
59	이탈	SOP20	2.000
60	이탈	SOP10	3.000



Set Globals

Settings Annotations

Globals to be created:

Field	MEAN	SUM	MIN	MAX	SDEV
고객등급	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Default operation(s): ☒ MEAN ☒ SUM ☒ MIN ☒ MAX ☒ SDEV

☒ Clear all globals before executing

☐ Display preview of globals created after execution

OK Run Cancel Apply Reset

Type

Preview

Types Format Annotations

Read Values Clear Values Clear All Values

Field	Measurement	Values	Missing	Check	Role
이탈여부	Flag	이탈/유지		None	Input
핸드셋	Nominal	ASAD170...		None	Input
고객등급	Continuous	[1.0,5.0]	*	None	Input

View current fields View unused field settings

OK Cancel Apply Reset

왼쪽 마우스 click!

On(*)선택

Missing

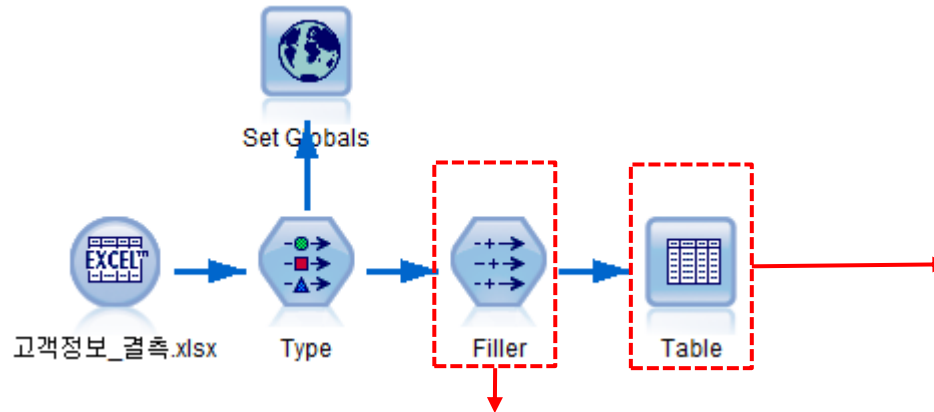
On ...

On (*)

Off

Specify...

결측 값 처리 (2/2)



대체된 결측값

Table (11 fields, 31,769 records) #15

	이탈여부	핸드셋	고객등급
41	이탈	SOP20	1.000
42	이탈	SOP10	4.000
43	이탈	SOP10	4.000
44	이탈	SOP20	1.000
45	이탈	SOP20	5.000
46	이탈	SOP10	3.000
47	이탈	SOP10	2.983
48	이탈	SOP10	5.000
49	이탈	SOP20	2.000
50	유지	SOP20	4.000
51	이탈	SOP20	5.000
52	이탈	SOP10	4.000
53	이탈	SOP20	3.000
54	이탈	SOP10	1.000
55	이탈	SOP10	2.000
56	이탈	SOP20	5.000
57	이탈	SOP10	2.983
58	이탈	SOP20	3.000
59	이탈	SOP20	2.000
60	이탈	SOP10	3.000

‘고객등급’
의 평균값
으로 결측
값을 대체
하는 대화
상자

Filler

Preview

Settings Annotations

Fill in fields:

고객등급

Replace: Based on condition

Condition:

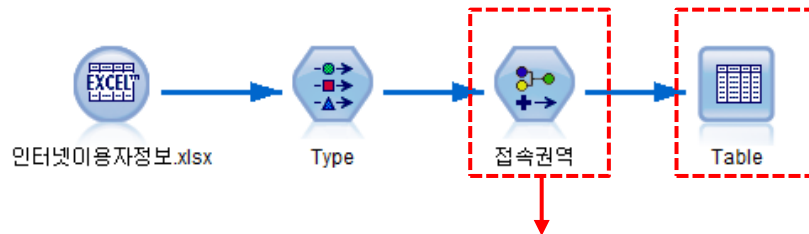
@BLANK(@FIELD)

Replace with:

@GLOBAL_MEAN(@FIELD)

OK Cancel Apply Reset

데이터 재분류



❖ 재분류(Reclassify) 노드

이산형 변수에 대해 여러 가지 범주 값을 하나의 범주 값으로 바꾸거나 기존 값을 다른 값으로 바꿀 때 사용

Mode: ☒ Single ☐ Multiple

Reclassify into: ☒ New field ☐ Existing field

Reclassify field: LOCATION

New field name: 접속권역

Reclassify values:

Original value	New value
강원	충청강원권
경기	수도권
경상	영남권
광주	호남권
대구	영남권

For unspecified values use: ☒ Original value ☐ Default value undef

	GENDER	JOB	LOCATION	접속권역
1	남자	무직/기타	충청	충청강원권
2	여자	방송/예술/스포츠	서울	수도권
3	여자	자영업	경기	수도권
4	남자	서비스	광주	호남권
5	남자	사무관리직	경기	수도권
6	남자	사무관리직	경상	영남권
7	여자	자영업	부산	영남권
8	남자	영업/마케팅	인천	수도권
9	남자	자영업	울산	영남권
10	남자	영업/마케팅	서울	수도권
11	남자	영업/마케팅	부산	영남권
12	남자	사무관리직	경상	영남권
13	남자	자영업	경기	수도권
14	여자	사무관리직	서울	수도권
15	여자	사무관리직	인천	수도권
16	여자	사무관리직	경기	수도권
17	남자	자영업	대전	충청강원권
18	남자	무직/기타	경기	수도권
19	남자	의료	부산	영남권
20	남자	영업/마케팅	부산	영남권
21	남자	사무관리직	서울	수도권
22	남자	서비스	경기	수도권
23	남자	교육직	부산	영남권
24	남자	서비스	서울	수도권
25	남자	영업/마케팅	인천	수도권

각 LOCATION을 수도권, 충청강원권, 영남권, 호남권의 대권역으로 입력

데이터 재구성 (1/3)

BEFORE

Table (5 fields, 70,274 records) #2

	CUS_ID	TIME_ID	CTG_NM	SITE_CNT	ST_TIME
1	1	2013061008	커뮤니티	1	46
2	1	2013031716	쇼핑	1	5
3	1	2012082902	인터넷/컴퓨터	2	53
4	1	2012081122	뉴스/미디어	1	0
5	1	2012122721	게임	2	0
6	1	2012081008	엔터테인먼트	1	0
7	1	2012092716	인터넷/컴퓨터	1	1
8	1	2012083017	엔터테인먼트	2	73
9	1	2012091813	인터넷/컴퓨터	2	6
10	1	2012080915	뉴스/미디어	1	136
11	1	2013060915	인터넷/컴퓨터	4	1
12	1	2012090506	인터넷/컴퓨터	4	10
13	1	2012081708	인터넷/컴퓨터	3	42
14	1	2012070918	인터넷/컴퓨터	38	107
15	1	2012071301	뉴스/미디어	1	67
16	1	2012112623	인터넷/컴퓨터	2	6
17	1	2012080507	커뮤니티	1	2
18	1	2012071322	인터넷/컴퓨터	11	86
19	1	2012112311	게임	1	0
20	1	2012083122	커뮤니티	1	14

Restructuring

AFTER

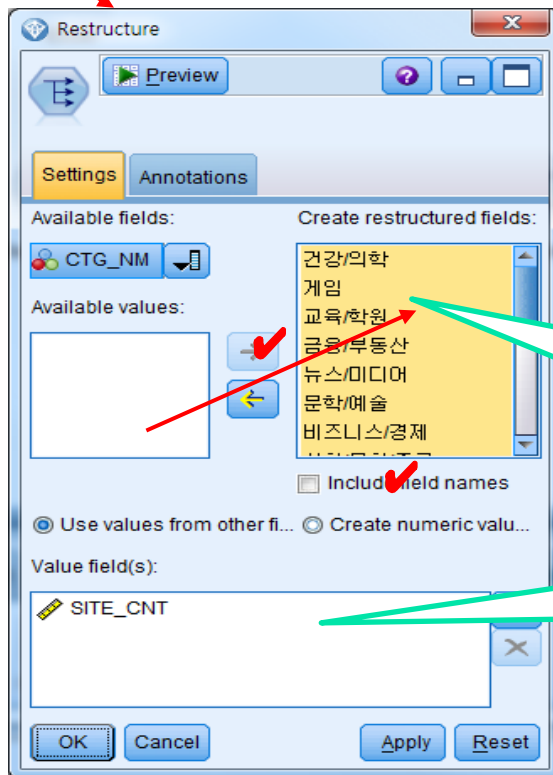
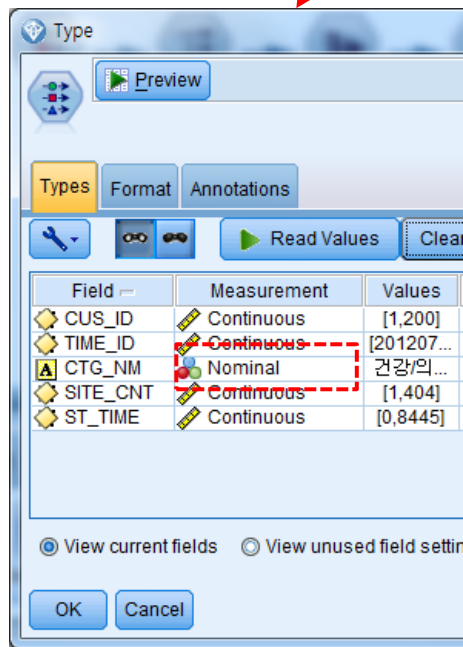
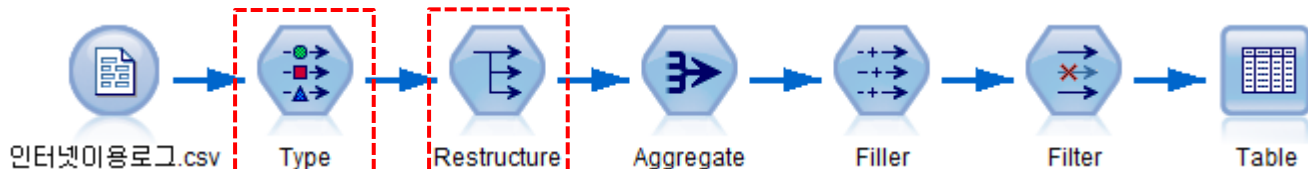
Table (23 fields, 200 records) #4

	CUS_ID	건강/의학	게임	교육/학원	금융/부동산	뉴스/미디어
1	1	0	36	0	0	258
2	2	0	0	0	15	83
3	3	8	0	0	3	153
4	4	1	0	0	35	215
5	5	3	2	0	51	495
6	6	0	0	0	16	1
7	7	2	0	0	160	248
8	8	0	0	0	66	11
9	9	1	303	0	0	146
10	10	1	396	0	50	171
11	11	0	4	0	45	8
12	12	0	56	0	14	2
13	13	0	0	0	199	292
14	14	0	0	5	9	0
15	15	0	0	0	55	13
16	16	0	0	0	0	10
17	17	0	0	0	31	40
18	18	0	0	4	362	87
19	19	14	1	0	325	87
20	20	0	0	0	77	8

데이터 재구성 (2/3)

❖ 재구성(Restructure) 노드

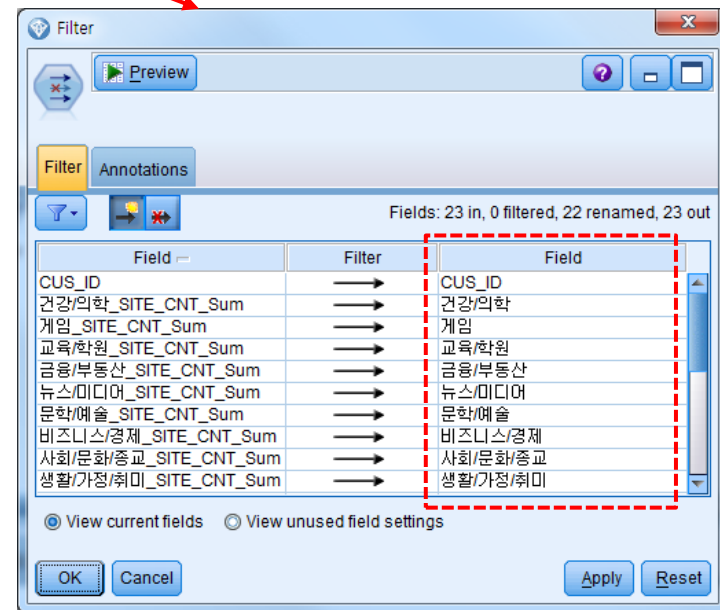
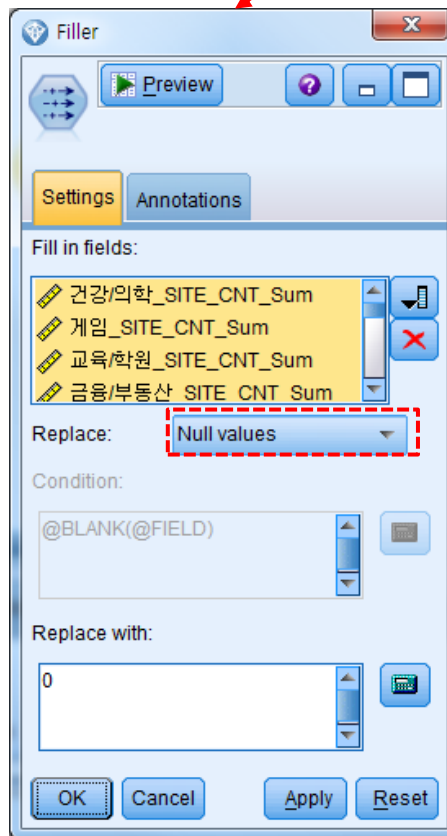
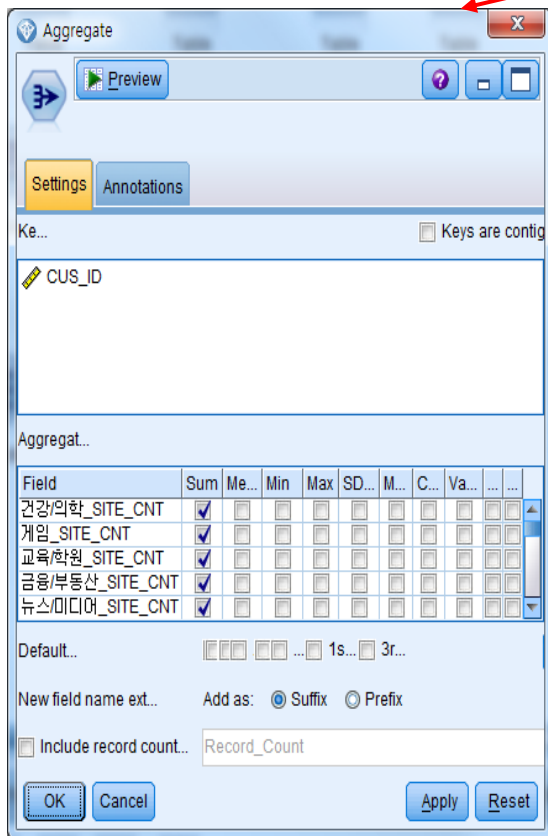
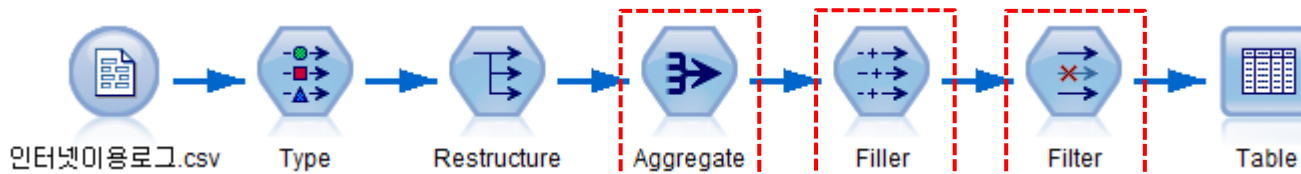
필드 유형이 nominal 또는 flag인 필드의 값을 이용하여 다중 필드를 생성(SetToFlag 노드의 일반형).



CTG_NM 필드의 값 -> 새로 생성할 필드명

SITE_CNT 필드의 값으로 새로 생성되는 필드의 값을 채움.

데이터 재구성 (3/3)



이상치(Outlier) 처리 (1/2)

인터넷이용자정보.xlsx → Type → [SCH_KEYWORDS] → Histogram with Normal Distribution

[SCH_KEYWORDS] → 10 Fields → Data Audit of [10 fields] #6

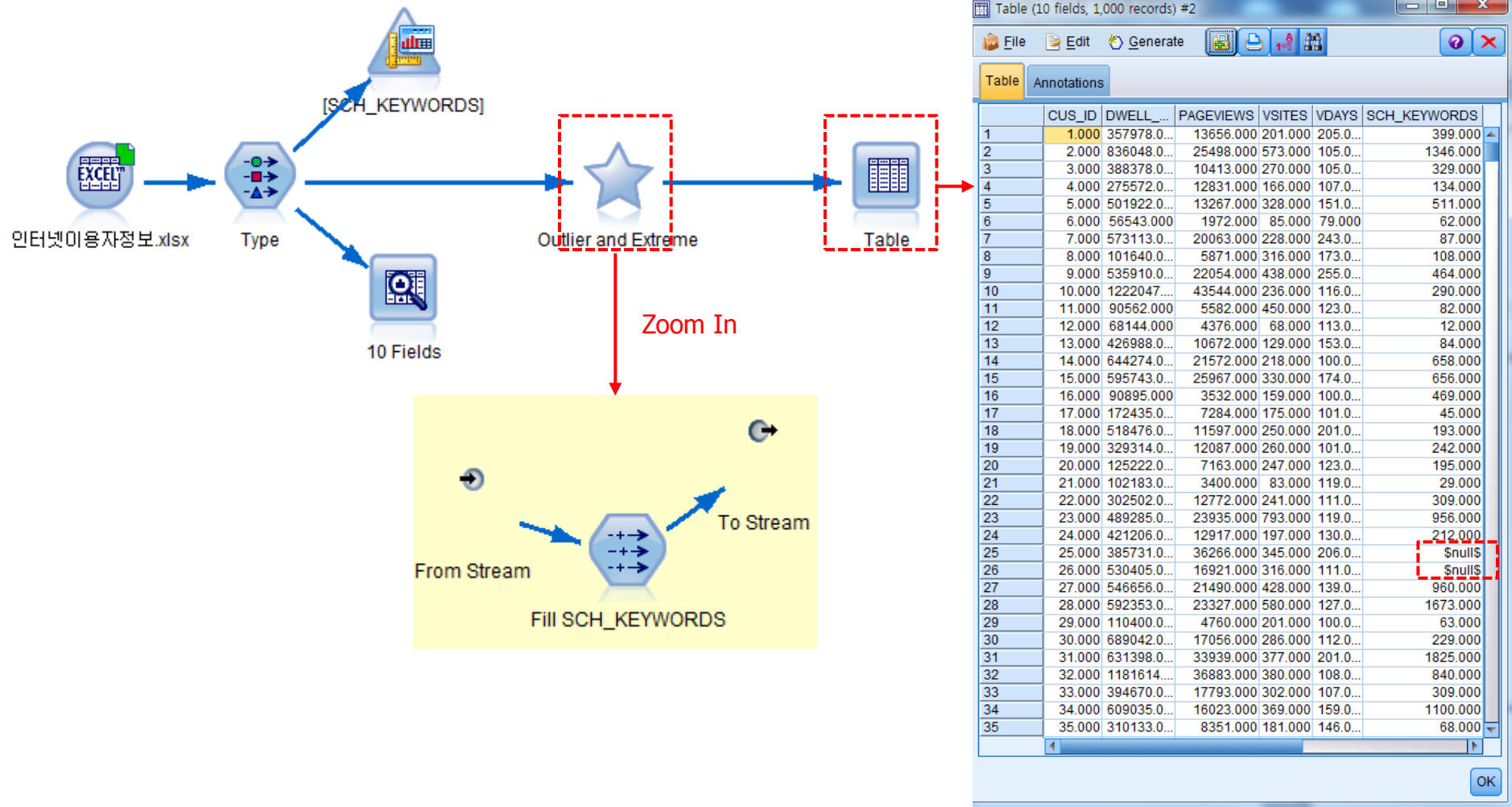
이상치 처리를 위한 슈퍼노드 생성

SCH_KEYWORDS 필드에 존재하는 21개의 이상치에 대한 처리방법을 결정

Missing Values SuperNode
Outlier & Extreme SuperNode
 Missing Values Filter Node
 Missing Values Select Node
 Reclassify Node
 Binning Node
 Derive Node

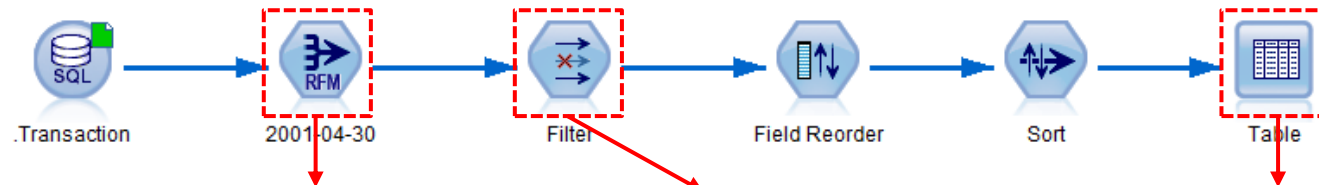
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method
CUS_ID	Continuous	0	0	None	Never	Fixed
DWELL_TM	Continuous	15	3	None	Never	Fixed
PAGEVIEWS	Continuous	8	1	None	Never	Fixed
VSITES	Continuous	13	1	None	Never	Fixed
VDAYS	Continuous	12	1	None	Never	Fixed
SCH_KEYW...	Continuous	19	2	None	Never	Fixed
BIRTH	Continuous	12	0	None	Never	Fixed
GENDER	Flag	--	--	Coerce	Never	Fixed
JOB	Nominal	--	--	Discard	Never	Fixed
LOCATION	Nominal	--	--	Discard	Never	Fixed
				Nullify		
				Coerce outliers / discard extremes		
				Coerce outliers / nullify extremes		

이상치(Outlier) 처리 (2/2)



파생변수 - 기간별 구매 금액·횟수·여부 (1/2)

- H백화점 고객의 최근 12개월 구매금액 및 구매횟수에 대한 변수 생성



2001-04-30

Preview

Settings Annotations

Calculate Recency relative to: ☒ Fixed date 2001-04-30 ☐ Today's date

☐ IDs are contiguous

ID: custid ✓

Date: sales_date ✓

Value: net_amt ✓

New field name extension: 최근12개월 Add as: ☐ Suffix ☒ Prefix

☐ Discard records with value below: 1.0

☒ Only include recent transactions. ✓

☐ Transaction date after: 2015-03-19

☒ Transaction within the last: 12 Months

☐ Save date of second most recent transaction

☐ Save date of third most recent transaction

OK Cancel Apply Reset

Filter

Preview

Filter Annotations

Fields: 4 in, 1 filtered, 2 renamed, 3 out

Field	Filter	Field
custid	→	custid
최근12개월Recency	→	최근12개월Recency
최근12개월Frequency	→	최근12개월_구매횟수
최근12개월Monetary	→	최근12개월_구매액

☒ View current fields ☐ View unused field settings

OK Cancel Apply Reset

Table (3 fields, 50,000 records) #2

File Edit Generate

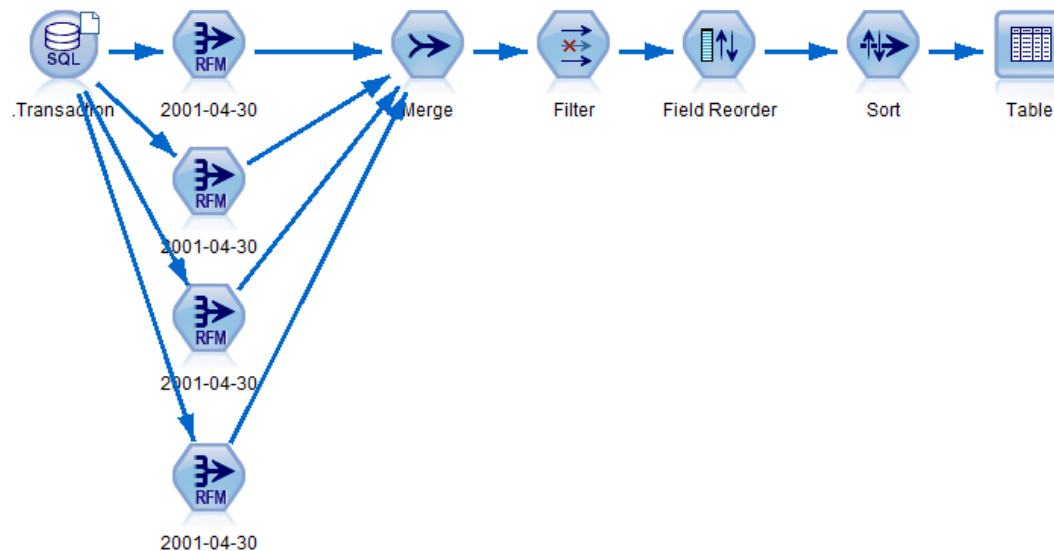
Table Annotations

	custid	최근12개월_구매횟수	최근12개월_구매액
1	1	77	3758381
2	2	28	2061179
3	3	68	6426049
4	4	6	381300
5	5	4	155700
6	6	29	2017400
7	7	11	2670000
8	8	244	10407549
9	9	28	1011680
10	10	42	1222175
11	11	12	653040
12	12	4	147820
13	13	11	2179300
14	14	37	5889350
15	15	2	83800
16	16	34	2736084
17	17	5	367350
18	18	28	1125200
19	19	5	320150
20	20	6	1060734

OK

파생변수 - 기간별 구매 금액·횟수·여부 (2/2)

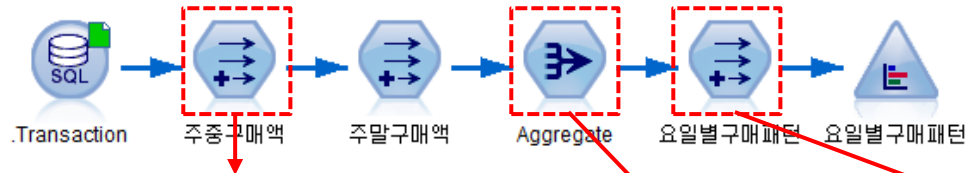
- H백화점 고객의 최근 1개월, 3개월, 6개월, 12개월 구매금액 및 구매 횟수에 대한 변수를 생성하시오.



- 구매액 또는 구매횟수 기준으로 볼 때 매출이 줄어드는 고객은 누구인가? 반대로 매출이 증가하고 있는 고객은?

파생변수 - 요일별 구매패턴

- H백화점 고객의 주중 • 주말 구매패턴에 대한 변수 생성



주중구매액

Derive as: Conditional

Settings Annotations

Mode: ☒ Single ☐ Multiple

Derive field:

주중구매액

Derive as: **Conditional**

Field type: **<Default>**

If:

`datetime_weekday(sales_date) = 1 or datetime_weekday(sales_date) = 7`

Then:

net_amt

Else:

0

OK Cancel Apply Reset

Aggregate

Derive as: Conditional

Settings Annotations

Keys are co

custid

Aggre...

Field	Sum	Me...	Min	Max	SD...	Me...	Co...	Va...	1st ...	3rd
주중구매액	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
주말구매액	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Def...

New field name ... Add as: ☒ Suffix ☐ Prefix

Include record co... Record_Count

OK Cancel Apply Reset

요일별구매패턴

Derive as: Nominal

Settings Annotations

Mode: ☒ Single ☐ Multiple

Derive field:

요일별구매패턴

Derive as: **Nominal**

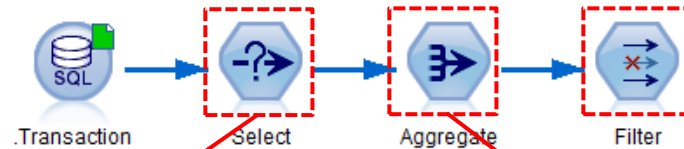
Field type: **Nominal** Default value: **없음**

Set field to	If this condition is true
주중	주중구매액_Sum >= 주말구매액_Sum * 1.5
주말	주말구매액_Sum >= 주중구매액_Sum * 1.5

OK Cancel Apply Reset

파생변수 - 환불행태 (1/2)

- H백화점 고객의 환불행태(금액, 건수)에 대한 변수 생성



Select

Preview

Settings Annotations

Mode: ☒ Include ☐ Discard

Condition:

net_amt<0

OK Cancel Apply Reset

Aggregate

Preview

Settings Annotations

Keys are c

custid

Aggr...

Field	Sum	Me...	Min	Max	SD...	Me...	Co...	Va...	1st ...	3r...
net_amt	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

D... ☐ 1s... ☐ 3r...

New field nam... Add as: ☒ Suffix ☐ Prefix

☒ Include record ... 환불건수

OK Cancel Apply Reset

Filter

Preview

Filter Annotations

Fields: 3 in, 0 filtered, 1 i

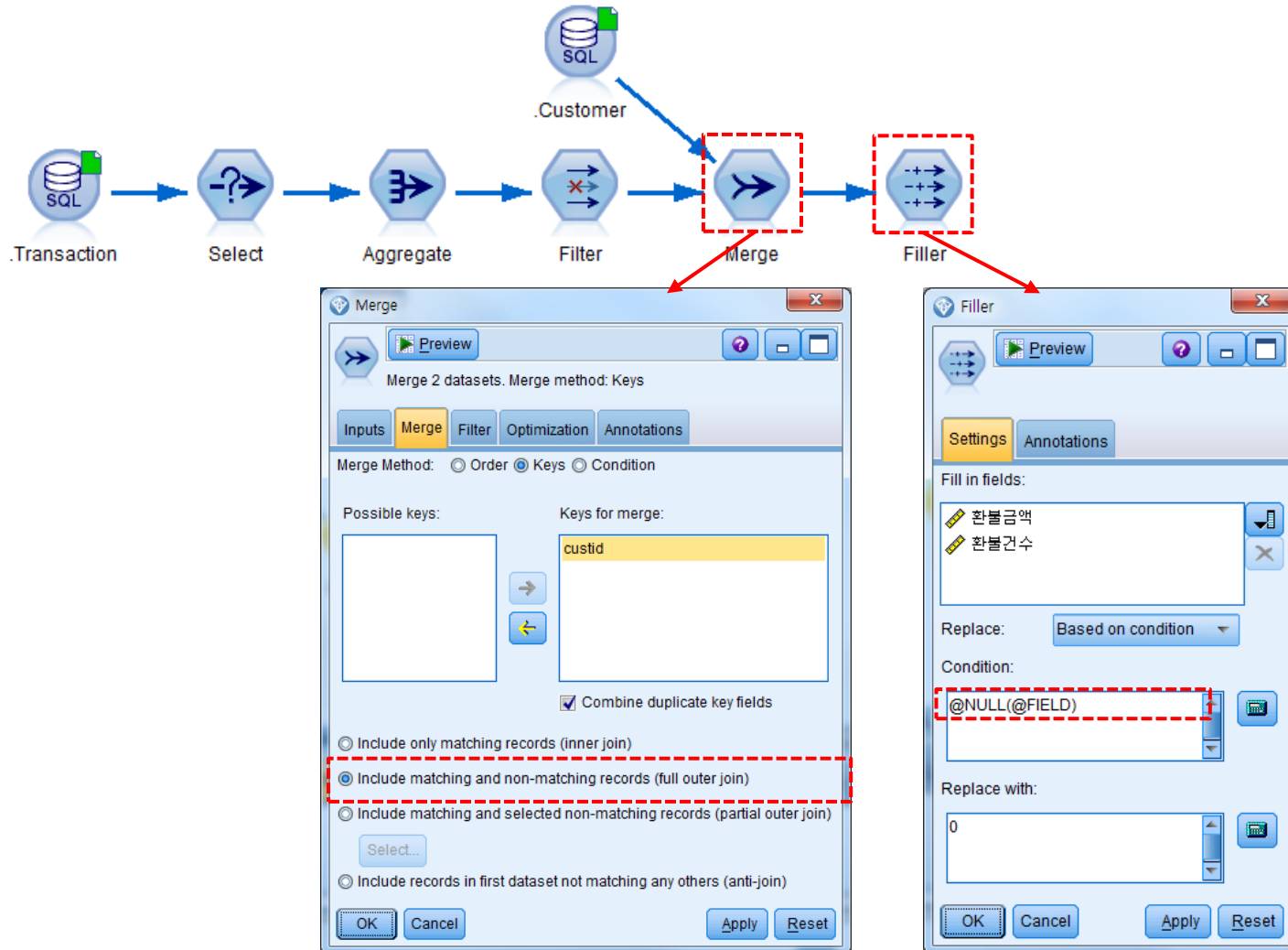
Field	Filter	Field
custid	→	custid
net_amt_Sum	→	환불금액
환불건수	→	환불건수

☒ View current fields ☐ View unuse...

OK Cancel Apply Reset

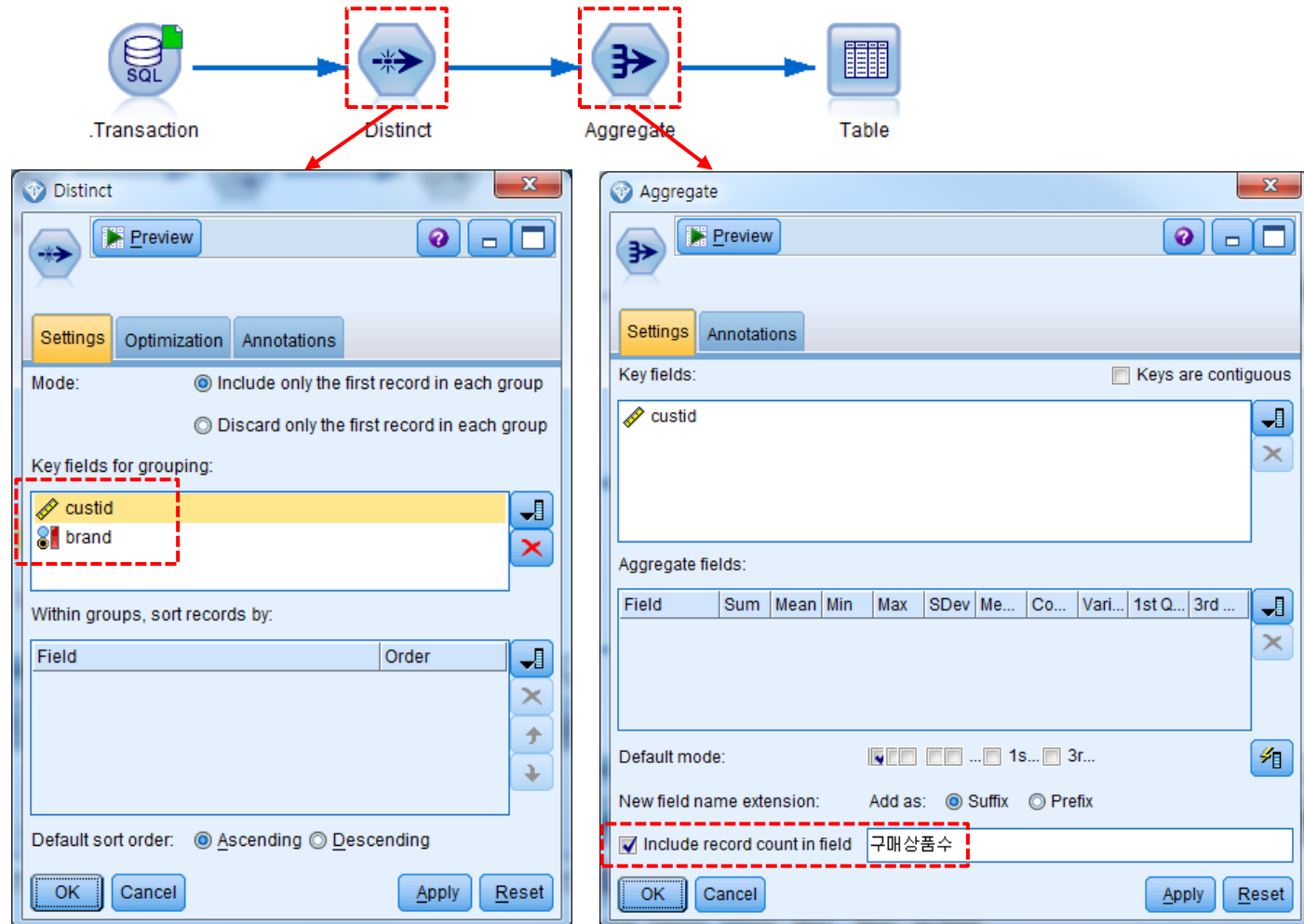
파생변수 - 환불행태 (2/2)

- H백화점 고객속성에 환불행태(금액, 건수) 변수를 결합



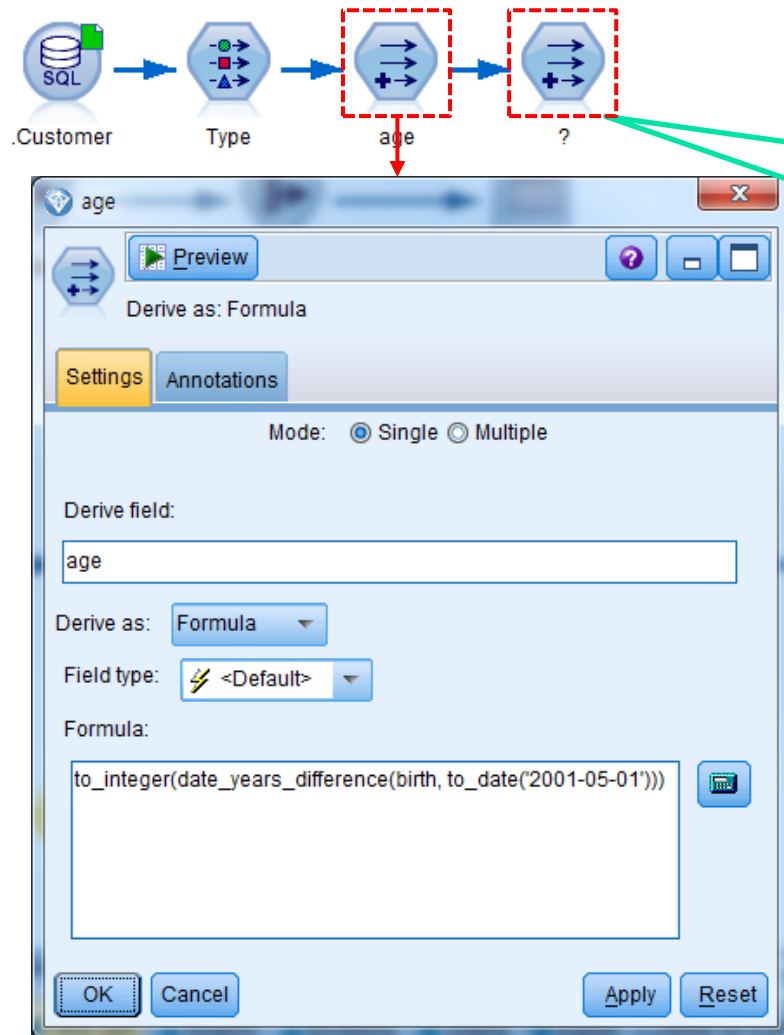
파생변수 - 구매상품 다양성

- H백화점 고객의 구매상품 다양성에 대한 변수 생성



파생변수 - 연령대

- H백화점 고객의 생일로부터 특정시점의 연령대를 계산하시오.

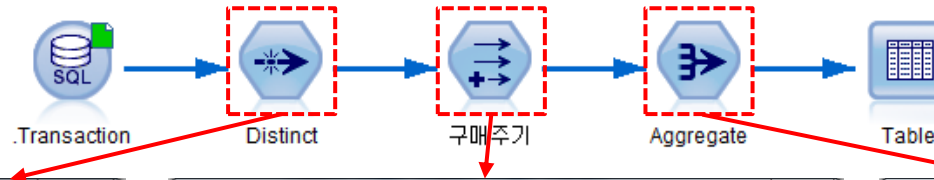


10대(이하), 20대, 30대, 40대, 50대, 60대, 70대(이상)으로 구간을 나눔

주의) 이상치 처리가 필요함.

파생변수 - 구매주기

- H백화점 고객의 평균구매주기(Average Purchasing Interval)를 계산



Distinct configuration window. The 'Settings' tab is active. Under 'Key fields...', 'custid' and 'sales_date' are listed and highlighted with a red dashed box. The 'Within groups, sort by' section is empty. The 'Default sort order' is set to 'Ascending'.

구매주기 configuration window. The 'Settings' tab is active. 'Derive as:' is set to 'Conditional' (marked with a red checkmark). 'Field type:' is set to '<Default>'. The 'If:' condition is '@OFFSET(custid,1) = @OFFSET(custid,0)'. The 'Then:' condition is 'date_days_difference(@OFFSET(sales_date,1),@OFFSET(sales_date,0))'. The 'Else:' condition is 'undef'.

Aggregate configuration window. The 'Settings' tab is active. 'Key field:' is 'custid'. 'Aggregate field:' is '구매주기'. The 'Default mode' is '1s...'. 'New field name extension' is 'Record_Count'. The 'Include record count in field name' checkbox is checked.



개인과제 #1 -6월5일까지 e메일제출

H백화점 데이터를 이용하여 아래와 같은 파생변수를 만드시오.

- 가격 선호도 변수
- 시즌 선호도 변수
- 상품별 구매 금액/횟수/여부 변수
- 상품별 구매순서 변수
- 주 구매상품 변수
- 휴면/이탈 가망 변수
 - Ex) If 평균구매주기 < 최종구매경과(현재 - 마지막 구매시점) Then “이탈”